

Weekly report

1 Done

1.1 Project

Related information can be found in another document.

1.2 Revision

相异性: (相似性=1-相异性)

-二元属性列联表

	1	0
1	q	r
0	s	t

-对称属性如性别

-对称二元属性: $(r+s)/(q+r+s+t)$

-非对称二元属性: $(r+s)/(q+r+s)$

-文档相似性: 构建词向量, 余弦相似

$$Z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Ordinal:

距离:

-欧几里得距离: 差的平方的和开根号

-曼哈顿距离: 差的绝对值的和

-闵可夫斯基距离: 差的绝对值的 h 次方的和开 h 次根号

数据处理:

-Cleaning, integration, reduction, transformation and discretization

-缺失值: 忽略元组、人工填写缺失值、使用全局常量或属性中心度量或同类数据均值或中位数或最有可能的值替换

-数据质量: accuracy, completeness, consistency, timeliness (随时间更新), believability, interpretability (易理解)

-采样原则: 选择有代表性的数据子集。Stratified sampling (分层采样) SRSWOR 无放回采样 SRSWR 有放回采样

-数据清洗: incomplete/missing data 设全局常量“unknown”; 默认平均值; 用贝叶斯或决策树推理

-数据集成: 从不同数据源整合数据元, 识别同一实体, 探测解决数据值冲突

-数据归约: smoothing; feature/attribute construction; aggregation; normalization; discretization

-z-score normalization: $v' = (v-\mu)/\sigma$

一些统计数据:

-Positively skewed data mode<median<mean

-Negatively skewed data mode>median>mean

-Quartiles: Q_1 25 分为点, Q_3 75 分为点

- Inter-quartile range: $IQR=Q_3-Q_1$
- Five number summary: min, Q_1 , median, Q_3 , max
- histograms often tell more than boxplots
- Quantile Plot 分位数图
- Q-Q Plot 两个分布比较
- 散点图：可以直观看到数据点集、轮廓灯
- 标称数据的卡方相关检验：Pearson χ^2 chi-square
(联合概率 o_{ij} 和期望频度 e_{ij} 差的平方除以 e_{ij}) 的和
 e_{ij} 是 $A=a_i$ 的数量乘以 $B=b_i$ 的数量除以总数
- 自由度一个属性的值的个数减一乘以另一个属性的值的个数减一
所得大于相应自由度和置信水平下拒绝假设的值，可以拒绝独立假设。
- 协相关系数 r =协方差除以两个标准差的积

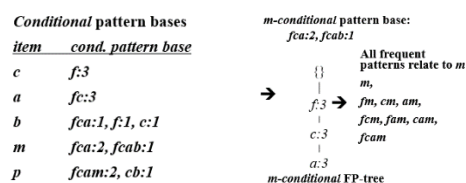
$$r_{A,B} = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^n (a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

- 分箱（去噪方法）：
用箱的均值或边界光滑

频繁模式:

- (absolute) support, or, support count of X: Frequency or occurrence of an itemset X
- (relative) support, s , is the fraction of transactions that contains X (i.e., the probability that a transaction contains X)
- X 是频繁的, 当它的支持度大于下限
- 支持度: 发生 X 和 Y 的概率
- 置信度: 发生 X 之后发生 Y 的概率
- Closed pattern, 没有和这个支持度一样且包含的 pattern
- Max pattern, 没有包含这个的频繁 pattern 了
- Apriori:
计算频繁一项集, 去掉支持度小于最小支持度的部分, 剩余里面找二项集, 计数, 去掉小于支持度部分, 以此类推。K-项集由 (k-1) 项集字典排序后, 前 k-2 项相同, k-1 项不同的链接
- Apriori pruning principle: 一个项集不频繁, 包含它的项集都不频繁
- Join step: k 项集的候选是从 k-1 项集中得出的
- FP 树:

支持度排序，根节点为空，第几层是第几个 条件模式基找前缀
conditional pattern base 条件模式基



分类:

- 有监督：训练数据有观察得到的类标签，新数据根据训练集分类
- 无监督：类标签不知道，建立类

-决策树:

分裂方法:

信息熵:

-ID3 信息增益: 偏向于多值的属性

$$\text{Info}(D) = I(Y, N) = -\sum p_i \log_2(p_i)$$

$$\text{期望信息 } \text{Info}_A(D) = \sum |D_j|/|D| * \text{Info}(D_j)$$

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D) \text{ 选最大}$$

-C4.5 信息增益率, 倾向于不平衡的划分

$$\text{SplitInfo}_A(D) = -\sum (|D_j|/|D|) * (\log_2(|D_j|/|D|))$$

$$\text{GainRatio}(A) = \text{Gain}(A) / \text{SplitInfo}(A)$$

选 GainRatio 最大的。

-CART 基尼指数, 偏向于多值属性, 类别大时难算, 倾向于在两个分区中产生相等大小的分区和纯度的分类

$$\text{Gini}(D) = 1 - \sum p_j^2$$

$$\text{Gini}_A(D) = |D_1|/|D| \text{gini}(D_1) + |D_2|/|D| \text{gini}(D_2)$$

$$\Delta \text{gini}(A) = \text{gini}(D) - \text{gini}_A(D)$$

这个值最小的节点被分离

分区 D 一开始是完全集, attribute_list 是描述元组属性的列表, attribute_selection_method 指属性的启发式过程。若 D 中的元组都在同一类 C 中返回 N 作为叶节点以类 C 标记。否则调用 attribute_selection_method 确定分裂准则, 指出分裂属性和分裂点。当 attribute_list 为空返回 N 作为叶节点, 标记 D 中的多数类。

停止条件: 分区 D 中所有元组属于同一个类; 没有属性可以用来进一步划分元组; 分区 D 为空。

剪枝以降低噪声和离群点带来的影响。先剪枝是提前停止树的构建, 后剪枝是删除结点。

贝叶斯:

后验概率 $P(H|X)$ 在知道 X 的基础上计算 H

先验概率 $P(H)$ 不基于其他信息

Naïve Bayesian

$P(X|H)$ 是已知个体符合 H, 则符合 X 的概率

$P(H|X) = P(X|H)P(H)/P(X)$ 。 $P(X)$ 为常数, 比剩下的部分, $P(H)$ 未知假设相等。多属性数据集, 假设与类条件相互独立 $P(X|C_i) = \prod P(x_k|C_i)$

A_k 是分类属性, $P(x_k|C_i) = C_i$ 中 x_i 数量 / C_i 数量

$$A_k \text{ 是连续属性, } g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

0 概率: 若 $P(x_k|C_i)$ 为 0, 则 x 也会高概率属于 C_i 但是 $P(X|C_i)$ 为 0, 假定训练数据库 D 很大, 对每个计数+1 避免 0 概率

优点: 好算, 大多数情况结果好

缺点: 假设条件导致失真

分类器性能评估 (TF 为分类正误, PN 为元组正负):

TP	FN	P
FP	TN	N
P'	N'	

准确率: $(TP+TN)/(P+N) = T/A$

错误率: $(FP+FN)/(P+N) = F/A$

敏感度: TP/P

特效性: TN/N

精度: $TP/(TP+FP) = TP/P$

召回: $TP/(TP+FN)$

$F_\beta = ((1+\beta^2) * \text{precision} * \text{recall}) / (\beta^2 * \text{precision} + \text{recall})$

F-score = F_1

ROC Curves: 纵轴 TP 横轴 FP 最优 area=1

选择因素: 精确度、速度 (构建和使用)、鲁棒性 (抗噪和缺失值)、可拓展性 (efficiency in disk-resident databases)、可解释性

检验: holdout 保持方法和随机二次抽样/k-折交叉验证, 分 k 个, 一次拿第 i 个当训练集, 错误率结果符合 k-1 个自由度的 t 分布, $|t| > z$ 可以拒绝均值相同假设

$$t = \frac{\overline{err}(M_1) - \overline{err}(M_2)}{\sqrt{\text{var}(M_1 - M_2)/k}}$$

TPR = TP/P FPR = FP/N = 1-specificity

ROC 曲线显示 TPR 和 FPR 之间的权衡

Cluster: 类间远类内近

-K 均值: $O(kn)$, k 均值只能用于连续值, 类别用众数, 需要确定 k, 对噪声和偏离点敏感, 不适合没有外壳的集合

任选初始中心, 分配对象至最近的中心, 根据族的位置重新计算均值为中心, 以此类推。对离群点敏感 $O(nkt)$ 可能局部最优。

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg \min_j \|x_n - \mu_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

$$\min_{r_{nk}, \mu_k} \sum_{\{n=1\}}^k \sum_{\{k=1\}}^k r_{nk} \|x_n - \mu_k\|$$

$$2 \sum_{n=1}^N r_{nk} (x_n - \mu_k) = 0$$

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}.$$

- K 中心点 (PAM): $O(k(n-k)^2)$ 鲁棒性较好, 也需要定义 k, 对小数据集有效, 大数据集计算复杂度过高

随机选对象代替中心点计算代价

支持向量机:

搜索最大边缘平面。

$w^T x_i + b \geq 1$ if $y_i = 1$

$$\mathbf{w}^T \mathbf{x}_i + b \leq -1 \quad \text{if } y_i = -1$$

$$y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1$$

每个点到分隔面的距离:

$$r = \frac{\mathbf{w}^T \mathbf{x} + b}{\|\mathbf{w}\|}$$

$$\min_{i=1, \dots, n} |\mathbf{w}^T \mathbf{x}_i + b| = 1$$

$$\rho = \frac{2}{\|\mathbf{w}\|}$$

所以中间的空 尽可能大

$$\mathcal{L}_P = \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^l \alpha_i - \sum_{i=1}^l \alpha_i y_i \mathbf{w}^T \mathbf{x}_i - b \sum_{i=1}^l \alpha_i y_i$$

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \quad b = y_k - \mathbf{w}^T \mathbf{x}_k$$

$$f(\mathbf{x}) = \sum_i \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b$$

线性回归:

$$f(\mathbf{x}) = a_0 + \sum_{i=1}^p a_i x_i = a_0 + \mathbf{a}^T \mathbf{x}$$

$$\hat{\mathbf{a}} = \operatorname{argmin} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \mathbf{a})^2 \text{ 要最小}$$

$$\min_{\mathbf{a}} \text{RSS}(\mathbf{a}) = (\mathbf{y} - \mathbf{X}\mathbf{a})^T (\mathbf{y} - \mathbf{X}\mathbf{a})$$

$$\frac{\partial \text{RSS}}{\partial \mathbf{a}} = -2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\mathbf{a})$$

$$\mathbf{X}^T (\mathbf{y} - \mathbf{X}\mathbf{a}) = 0$$

$$\text{假设 } \mathbf{X}^T \mathbf{X} \text{ 非奇 } \hat{\mathbf{a}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{a}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Ridge Regression:

$$\hat{\mathbf{a}}^{\text{ridge}} = \operatorname{argmin} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \mathbf{a})^2 + \lambda \sum_{j=1}^p a_j^2$$

$$\hat{\mathbf{a}}^{\text{ridge}} = \operatorname{argmin} [(\mathbf{y} - \mathbf{X}\mathbf{a})^T (\mathbf{y} - \mathbf{X}\mathbf{a}) + \lambda \mathbf{a}^T \mathbf{a}]$$

$$\hat{\mathbf{a}}^{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

Lasso:

$$\hat{\mathbf{a}}^{\text{lasso}} = \operatorname{argmin} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \mathbf{a})^2$$

$$\text{s.t. } \sum_{i=1}^p |a_i| < t$$

$$\hat{\mathbf{a}}^{\text{lasso}} = \operatorname{argmin} \left\{ \frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \mathbf{a})^2 + \lambda \right\}$$